MONTHLY WEATHER REVIEW

CHARLES F. BROOKS, Editor.

VOL. 49, No. 3. W. B. No. 736. | MARCH, 1921. | CLOSED MAY 3, 1921 ISSUED JUNE 4, 1921

THEORY AND USE OF THE PERIODOCRITE.[1]*

By CHARLES F. MARVIN, Chief, U. S. Weather Bureau.

[Washington, D. C., Mar. 1, 1921.]

SYNOPSIS.

Periodocrite [1] is a word coined from Greek roots signifying a critic, a judge, a decider of periodicities, and is a name applied to a mathematical and graphic method or device which has been developed to aid in the conclusive separation of obscure and hidden cycles and periodicities possessing a real existence from those whose essential features are only such as would result from, and can be explained by, entirely chance combinations of the data employed.

The periodocrite does not disclose or discover the *length* of suspected periods or cycles. Other methods, such as the harmonic analysis, Schuster's periodogram, or any of the many methods which have been offered for this purpose must first be employed to ascertain the proper length of any suspected cycle.

Suggestions for clearness in terminology are offered; the elements of the theory are briefly presented, the significance of the results secured in applications to practical data are illustrated by examples, and methods of abridging certain mathematical computations are indicated.

The paper concludes with representations concerning the inherent characteristics of data, the limitations upon the use of the familiar least square methods in problems of meteorology, and some suggestions are offered for overcoming difficulties thus entailed.

INTRODUCTION AND TERMINOLOGY.

How can claims of periodicities in the succession of values of meteorological and like data be proved or disproved?

If a definite and conclusive answer to this important question were known, vast expenditures of time and labor on the part of many students might have been spared in the past and like efforts conserved in the future for more fruitful pursuits than one so alluring and baffling as the search for obscure and hidden cycles believed by many students to lie concealed in practically any body of meteorological and like data.

Science must furnish conclusive answers to the claims and questions with which students of periodicities are confronted and this paper is an effort to supply a few basic principles which may serve as part of a greatly needed solid foundation upon which any superstructure of periodic theories and claims may be successfully erected.

If what is offered does not suffice fully to segregate specious sequences from cyclical or periodic elements which have a physical *raison d'être,* nevertheless it may guide the reader in forming his own conclusions and aid the investigator to avoid wasted effort on studies which must necessarily lead only to fruitless or inconclusive results.

Our acceptance and approval of claims of periodicities which may be advanced is justified only when it has been demonstrated that the propositions involved can successfully run the gauntlet of tests and criteria such as presented herein.

As if to add the last acid test to whatever may emerge as real, rather than fortuitous, from all other tests, we must invoke the quality of *utility*. The value of a scientific truth is, of course, wholly independent of any numerical measure on which the fact depends. The parallax of a remote star may be vanishingly small but of vast importance. On the other hand, a periodicity which differs from a perfectly fortuitous sequence by only a very small margin may be an important scientific truth to demonstrate, but its practical value for forecasting or other purposes may be entirely inconsequential. I have indicated in an unpublished note how the forecasting value of knowledge of this character can be measured.

The ultimate goal of the student who seeks to formulate the laws of sequences of solar and terrestrial phenomena and correlations thereof must be to establish, first, the REALITY, second, the UTILITY of claims.

For the present purposes, statistical data may be put into two classes:

Class I is illustrated by observations which exhibit the diurnal and annual march of temperature and pressure, the seasonal variations in rainfall, the 11-year period in sunspots, and like features in terrestrial magnetism. Simple inspection of such data or graphs of it leaves no doubt as to its periodic features. Either aided or not by known laws defining the length of the periods, the major question in these cases is to fix the amplitude and elements of the period, or, in the more complex case presented by sunspots and magnetic phenomena, to ascertain the real length of the changeable period and the nature of its fluctuations.

Class II comprises many other cases in which no real period can be discerned by simple inspection, and this will be true also in cases of the data constituting Class I after steps have been taken, as is often done, to free the original material of any diurnal, annual, or other periodic features which can be evaluated with greater or less accuracy.

Terminology.—Differences unwittingly and unintentionally put upon the meaning of terms and language by the different parties discussing a given subject is too often the only ultimate basis for material differences of view. No other department of meteorology is more subject to this possibility than the question of cycles and periodicities. This is because the precise and exact terms and language of physical and mathematical harmonics are sometimes carelessly applied to features of data which crudely recur, when in fact the terms employed are quite inapplicable unless they carry different meanings from those ordinarily signified.

It is most important that the meaning of the language and terms of mathematical harmonics be kept pure and significant of certain definite things and not loosely applied to features of meteorological data which have only the slightest resemblance to a cycle, a period, a harmonic, or what not, as the case may be. It is needless to define here the mathematical terms of periodicities.

*Presented before American Meteorological Society, Washington, April 20, 1921.

[1] Prof. C. F. Talman supplied this name from περίοδος, a period + κριτής a judge, decider, umpire, from κρίνω to separate, investigate, judge.

This is abundantly done in the textbooks, and when used in a meteorological connection those terms should carry the customary meaning as exactly as possible. If this be done in a legitimate way, then readers will have a basis for the clear understanding of authors. Many of the recurrent features of weather phenomena which writers call or claim are periods, cycles, and the like, are little more than indefinite "*sequences*."

The latter term seems particularly appropriate for meteorological application, and is offered for general use, defined somewhat as follows:

Sequence.—Any more or less complex succession of values of meteorological or other states, elements, etc., especially those which exhibit a tendency to cyclical or periodic recurrence. A portion of such sequences or a definite result derived therefrom which exhibits *marked periodic recurrence* of essential features may be properly called a *cycle* if somewhat complex, but when the form is definitely periodic and very simple, with essentially one maximum and one minimum value in the sequence, then the appropriate definitive designation, *periodic element*, is suggested. *Harmonic elements* or *harmonic components* are terms which should be used consistently to designate only those elements the inherent characteristics of which are sinusoidal, whereas periodic elements or components of weather sequences very often or nearly always are distinctly nonsinusoidal or nontrigonometric in any form, and the terminology should keep this inherent distinction clear.

The problem of the meteorologist and the forecaster seeking to extend the period of his forecasts far into the future, is to discover and formulate the *laws governing meteorological sequences*, if indeed those laws do not turn out to be indistinguishable from the laws of chance. This statement contains a very important truth deserving more recognition than it has received. Phenomena of chance are theoretically considered to be the outcome of the operation of a large number of independent influences or causes. Sequences of weather phenomena are also consequent to the operation of a large number of conspiring causes which in many particulars are quite independent, resulting in consecutive values whose laws of succession can be segregated from the laws of chance only with great difficulty and incompleteness. Progress in these studies has been hindered and delayed, no doubt, by the too hasty or too confident conclusion of many students that weather sequences can be *resolved* into cycles or *analyzed* into harmonic Fourier elements. The physical existence of an obscure cyclical component of data should be unequivocally proven before its reality is claimed, because errors of science advocated and diffused often on high authority tend to perpetuate themselves indefinitely, and their subsequent correction is exceedingly slow and difficult.

The Fourier series can REPRESENT *any succession of* variable values. It is utterly futile, however, in the great majority of cases, to push the application to more than an imperfect *representation*. Hourly values of temperature, for example, form a very nearly perfect periodic element. The sequence, however, is entirely nonharmonic. Its features are the outcome of chiefly two entirely independent processes: (1) An uninterrupted process of cooling which is subject to a large number of modifying influences such as cloudiness, winds, temperature, and nature of surface, etc. Cooling is always a losing operation, and this tendency to change as an elemental effect is always in one direction and is therefore absolutely nonharmonic, even nonperiodic. (2) The other control on diurnal temperature is the intermittent influence of solar heating, which begins at sunrise and is cut off at sunset, attaining various intensities at intermediate hours, depending on atmospheric transmissibility, etc. As an elemental effect, this also is absolutely nonharmonic because it is intermittent.

Of course other factors modify the hourly values of temperature, such as importation of warm or cold air from a distance, etc., but the diurnal march of temperature is cited as an excellent example of a *periodic element* in weather sequences which is entirely nonharmonic in its physical character. Accordingly, the harmonic analysis applied to such data is altogether meaningless except that the sum of the harmonic elements simply *represents* the original data.

On the other hand, the sequence of daily, weekly, or monthly values of mean temperature for a station or locality form a good example of an annual cycle whose features are found by both theory and observation to conform satisfactorily to a limited number of Fourier elements.

Sequences in values of pressure, humidity, cloudiness, precipitation, the departures of elements from average, and so-called normal conditions all quickly become very complex and nonharmonic. Resort to the use of devices and terminology of the harmonic analysis in the discussion of such data is more apt to mislead the student and confuse the reader than to disclose useful meteorological laws and principles.

It has seemed necessary to offer the foregoing discussion on terminology because the literature of the subject is often lacking in consistency and clearness, and the making of the distinctions and discriminations mentioned is necessary to an understanding of what follows

THE PERIODOCRITE.

It is assumed we have at hand a large number, N, o homogeneous values of any variant which is suspected to be characterized by one or more hidden cycles or periodic elements. The data may represent temperature, rainfall, sunspots, intensities of radiation, or what not, and it is assumed all *obvious* periodicities like the diurnal, annual, and other cycles have been eliminated by appropriate methods, also that the data have otherwise been brought into a homogeneous body of values of equal weight. Finally, it is assumed that a group or succession of p values of these data comprise a more or less complex cycle or periodicity which repeats itself over and over again. However, owing to the large accidental variations of the values of the variant, the cycle is hidden and can not be satisfactorily discerned from a careful inspection of the data.

The well-known method of evaluating or proving a periodicity in such a case consists in tabulating the data in rows and columns in a manner designed to *bring into the same columns data of the same suspected phase relations.* The accidental variations will thereby tend to average out, and the real features of the hidden periodicity will be exhibited by the sums and means of the phase columns. Emphasis is placed on the words in italics because in carrying out a tabulation in rows and columns it is equally possible to place in each of the several columns the same phase values of the data *when the length of the period is variable as when it is constant.* Attention to this point is necessary if we are to deal properly with periodicities, the length of which vary systematically and progressively as claimed by some investigators. The only difficulty introduced by the variability in the length of the period is the added labor of computation it entails.

Let a tabulation of any portion or all of a given body of data in p phase columns and n rows be indicated by letters as below, and for brevity let any group like this be designated a "*tabulation*":

A periodicity tabulation.

Number of cycles (n).	Phase values.							
	0	1	2	3	......	......	......	$p-1$
1	a_0	a_1	a_2	a_3	......	......	......	a_{p-1}
2	b_0	b_1	b_2	b_3	......	......	......	b_{p-1}
3	c_0	c_1	c_2	c_3	......	......	......	c_{p-1}
....	......	......	......	......	......	......	......	......
n	n_0	n_1	n_2	n_3	......	......	......	n_{p-1}
Sums	S_0	S_1	S_2	S_3	......	......	......	S_{p-1}
Means	m_0	m_1	m_2	m_3	......	......	......	m_{p-1}

The whole body of homogeneous data available is supposed to be indefinitely large, permitting of one or several relatively large "*tabulations*" of independent data.

Now, only one of two results is possible with reference to the sums S or the means m derived from adequate data.

(1) The sums, as likewise the means, will be constants—that is, differences will be small and negligible, and therefore no period is indicated, or

(2) The differences in the successive values of S or m will be of material magnitude and may possibly signify a cycle.

Experience with meteorological and like data teaches us that when n in a *tabulation* is very small, say 3 or 4, the variations in the values of m will be very large. In general, a mean of 4 sets of values will show only about one-half, and of 9 sets often only about one-third as much range of variation as exhibited by any one of the single sets. It is also generally found that when all obvious periodicities have been eliminated, the means of a *tabulation* show less and less differences among themselves as the value of n increases. These are relations which we know correspond entirely to the requirements of chance.

Accordingly, the question at once arises: Are the variations found in the values of m_0, m_1, m_2, etc., in any limited *tabulation* in any degree different from variations which would be found in values computed from entirely chance combinations of the same body of data? We may assume, for example, that all the values constituting the original observations are completely mixed up in a bowl and that any specified *tabulation* is made up entirely from pure chance drawings. The body of data thus secured will differ from the actual observation in only one particular, namely, *the order of succession* of the values. In all other features, as the mean, the mode, the median, the standard deviation, etc., the whole distribution of the chance drawings will be identical with like elements of the original observations which will differ only in the order of succession of values. We may well ask, therefore, how will values of m_0, m_1, etc., for actual observations differ from similar values derived entirely from chance drawings from the same body of numbers?

The search for the answer to this question led to the development of the periodocrite. It is a method and a graphic device which serves not only to segregate real from accidental periodicities, but with an adequate amount of data will afford a very satisfactory index number which shows the *realness* of the periodicity.

Theory.—The theory of the periodocrite is briefly developed as follows:

Let $\sigma_0 = \pm\sqrt{\frac{\Sigma V^2_N}{N}}$ = the standard deviation of the whole body of data, in which ΣV^2_N designates the sum of the squares of the departures, and N the total number of observations. The subscripts attached here and elsewhere to the symbol V^2 designate the particular group of data from which the sum of the squares of the departures is derived, in this case the whole body of data N. It is assumed σ is computed by forming the frequency distribution in the usual manner, and if the distribution is distinctly unsymmetrical, that fact should be ascertained and duly considered, together with any features of abnormality which may affect the data and which probably can not be easily removed.

Perfect fortuity.—We shall first consider the case of perfect fortuity. We must necessarily assume that in any *tabulation* of a portion of the whole data, the values in the portion, n rows and p columns, for example, or np values in all, are representative of the whole body of data. Of course the failure to satisfy this requirement always occurs in problems of chance, and it simply causes minor deviations from theory which are generally recognized and understood. On the assumption made, then, in the long run we may write:

$$\sigma_{np} = \pm\sqrt{\frac{\Sigma V^2_{np}}{np}} = \pm\sqrt{\frac{np\frac{\Sigma V^2_N}{N}}{np}} = \sigma_0$$

Also, if σ_n is the standard deviation of the p mean phase values of a *tabulation*, then

$$\sigma_n = \pm\sqrt{\frac{p\Sigma V^2_m}{np}} = \pm\sqrt{\frac{\Sigma V^2_s}{np}} = \pm\sqrt{\frac{\Sigma V^2_m}{n}}$$

Now if chance is the only factor which controls the results brought out in a *tabulation*, then from the principles of least squares we must have

$$\sigma_n = \frac{\sigma_0}{\sqrt{n}} \text{ from which } \frac{\sigma_n}{\sigma_0} = \frac{1}{\sqrt{n}} \qquad \text{--------- (1)}$$

Let y = the ratio $\frac{\sigma_n}{\sigma_0}$ which we may regard as a coefficient of variation.

Also let $\frac{1}{\sqrt{n}} = x$. Hence from (1)

$$y = x \qquad \text{-------- (2),}$$

which is the equation of a straight line through the origin of coordinates at an angle of 45° to the axes. From the derivation of its equation such a line represents the results of perfectly fortuitous combinations of the data employed. (See fig. 1.)

Since n is any integer from 1 to + infinity, the values of x lie between $x=0$ and $x=+1$. Likewise, the values of y range between $y=0$, and $y=+1$, because for perfectly fortuitous control when $n=+$infinity $\sigma_n=0$, hence $y=0$ and when $n=1$, σ_n *on the average* $=\sigma_0$ $\therefore$ $y=1$.

The application of the foregoing to actual observational data is very simple. Having computed the standard deviation σ_0 for all the data, find mean phase values m_0, m_1 ------ for one or several tabulations. Then deduce values of x and y as explained. If these are equal, or nearly so, then the variations in the values of m_0, m_1, m_2, etc., are no greater than those due entirely to chance, and at least in so far as the *amplitude* of variations is concerned no claims of periodicity are justified.

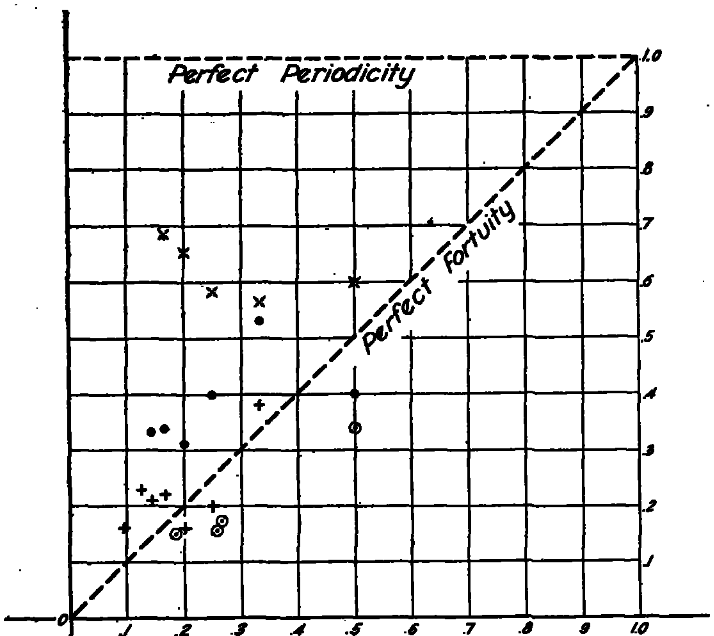

FIG. 1.—Rainfall periodocrite. Star: Annual cycle five stations in Iowa, 36-year record. Heavy dots: Annual cycle Washington rainfall, 50-year record. Crosses: Annual cycle Boston, Mass., 103-year record, very feebly defined. Circles: A 15-month sequence, Iowa rainfall. Other sequences, 15 months, 16 months, one-ninth the variable sunspot period, like the circles, all fall in the class of perfect fortuity.

Perfect periodicity.—Let us consider data which exhibits perfectly cyclical succession of values, as for example consecutive values of the average hourly temperatures repeating the sequence over and over again. Of course the periodic feature would be perfectly obvious in such a case and there would be no need to resort to analytical demonstrations to establish a periodicity, however the case aids in developing the theory of the periodocrite.

It is known that σ_0 is entirely independent of any question of the order of succession of the data. However, if p values in sequence constitute a complete cycle, then because the cycle is perfect a *tabulation* of np observations will give the same phase values of m_0, m_1, m_2, etc., whatever the value of n $\therefore$ $\sigma_n = \text{constant} = \sigma_0$ for perfect periodicity and $\therefore$ $y = \frac{\sigma_0}{\sigma_0} = \text{constant} = 1$ -------- (3), which is the equation of a line parallel to the axis of X at distance 1 and is a line of perfect periodicity.

From the foregoing we see in general that to test a periodicity it suffices to form one or more *tabulations* as explained and compute values of x and y for one or more groups of data. When y is substantially and consistently greater than x a real periodicity is indicated of greater or less amplitude. If x and y are nearly equal, especially y smaller than x, the amplitude of the periodic variations is less or no greater than that due wholly to chance. In the face of such a result, the probability of the cycle being real is very small or nil. An entirely new body of data may give a like range of values of phases, but the order of succession may be quite different. If, however, the order and features of succession of the phase values should prove to remain sensibly invariable even while x and y remain nearly equal for different *tabulations* of independent data, then the conclusion must be that a periodicity exists of amplitude no greater than chance alone will produce, and the period tends to vanish as the length of record increases.

It will be noticed that theoretically y can not exceed unity; however, when n is large y may become several times larger than x. This shows strongly marked periodicity, but in all such cases inspection alone establishes the same fact, and the calculation of x and y simply serve to express in numerical terms the proportionate effects of periodic control as compared with chance represented by x.

SUMMARY OF APPLICATIONS.

The principle of the periodocrite has been applied to a short study of rainfall in annual and approximate 15-month cycles, both fixed and variable in length. The results are shown partly in figure 2. Only the annual cycles show any real existence. Even this is very feeble at Boston, Mass., for a long record of 103 years. All other cases examined show a variation over the cycle just such as the laws of chance would lead us to expect, that is no period except the annual one has been found.

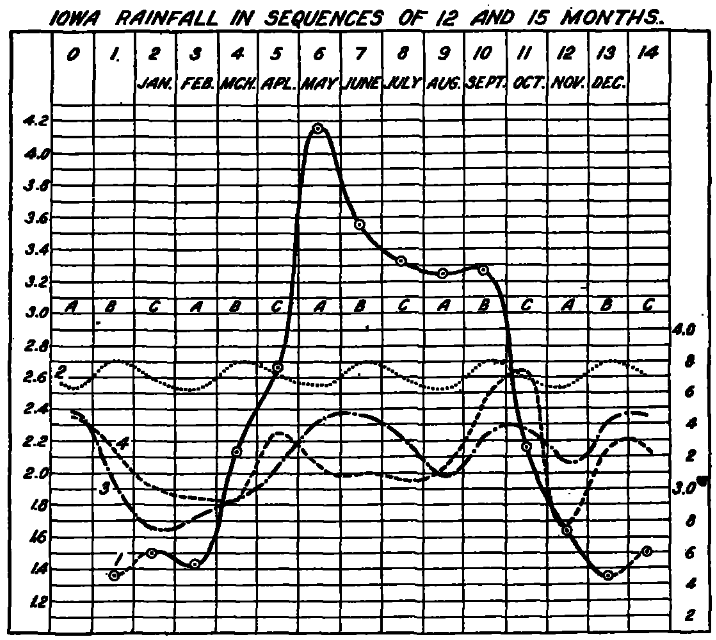

FIG. 2.—(1) Annual cycle, 36 years. (2) Effect of distributing the annual cycle through a 15-months sequence. (3) Actual 15-month cycle after eliminating annual cycle by ratios. (4) Same cycle, secured by fortuitous drawings from 36 years of monthly ratios.

ABRIDGED METHODS OF COMPUTATION.

Investigations of periodicities by the methods indicated in the foregoing require many calculations of σ, the standard deviation of the data. The quantity σ appears in the theory simply as a measure or index of the diversity or amount of variation either in the whole body of data or of certain portions thereof supposed to constitute a cycle or periodic element. Some other measure of variation, such for example as the average departure from the mean taken without regard to sign, will serve the same purpose and with nearly the same accuracy.[2] In fact, this index of diversity, $V_n = \frac{\Sigma V}{n}$ of any group of values, as of S_0, S_1, S_2, etc., or of m_0, m_1, m_2, etc in a *periodicity tabulation*, is quite as dependable for present purposes as $\sigma_n = \pm\sqrt{\frac{\Sigma V^2}{n}}$ or a value of the probab levariation, $E_n = \pm .6745\sqrt{\frac{\Sigma V^2}{n-1}}$ because in any one of these cases

[2] Davenport: C. B. Statistical methods with reference to biological investigations. New York, 1914. p. 16.

we are compelled to measure the variation of relatively small groups of numbers. As a random sample of the whole body of data, such groups are not representative. As a general rule, a small portion of the data will tend to show *less variation* than the whole body of data, because the very large departures which are due to occur infrequently are not likely to be found in a small sample taken at random. On the other hand, when such extremes do occur the variation for the small group then appears to be too great. All these considerations fully justify using simply $V_n = \pm\frac{\Sigma V}{n}$ as a measure of variation. The calculation of $\frac{\Sigma V}{n}$ is much easier than the quantity $\frac{\Sigma V^2}{n}$ and there is indicated below a quick method of computing ΣV without even forming the individual values of departures V. In this connection it is shown in the textbooks on least squares [3] that there is a direct relation between ΣV and the standard deviation σ. Thus it is shown that the probable error of a single observation is

$$r = \frac{.8453\ \Sigma V}{\sqrt{n(n-1)}}$$

From this it is easy to show that

$$\frac{\Sigma V}{n} = .7979\ \sigma_n \tag{4}$$

This relation is of much practical importance and is used in the present connection in the following manner:

To enter upon any rational study of any independent body of data a first important step is to reduce the data to as homogeneous and as elemental a state of values of equal weight as possible. Then form a frequency classification and compute the standard deviation. We designate this value

$$\sigma_0 = \pm\sqrt{\frac{\Sigma V^2_N}{N}}$$

From (4) we easily obtain

$$V_0 = \frac{\Sigma V_N}{N} = .7979\ \sigma_0$$

It is seen, therefore, to be good practice to compute the standard deviation directly from the *whole body of data* and from this to evaluate $\frac{\Sigma V}{N}$ indirectly, because this value is subsequently needed for comparative purpose with other values of $\frac{\Sigma V}{n}$ derived by abridged methods of computation, as follows:

Let $a_1, a_2, a_3, a_4 \ldots a_n$ be any group of values for which the arithmetical sum of the departures from the mean is desired.

Tabulating the usual operation of effecting such a calculation we get a table as follows in which the t positive and u negative departures are set out in different columns in which b_1, b_4, etc., indicates values of a which are greater than the mean and c_2, c_3, etc., values of a equal to or less than the mean.

[3] M. Merriman: Method of Least Squares. John Wiley & Sons. 1915. P. 92.

Tabulation of departures from mean.

Values.	Departures.	
	Positive.	Negative.
a_1	$b_1 - M \mp \frac{r}{n}$.	
a_2		$c_2 - M \mp \frac{r}{n}$.
a_3		$c_3 - M \mp \frac{r}{n}$.
a_4	$b_4 - M \mp \frac{r}{n}$.	
a_n		$c_n - M \mp \frac{r}{n}$.
Sums Σa	$\Sigma b - tM \mp \frac{r}{n}t$.	$\Sigma c - uM \mp \frac{r}{n}u$.

Mean $= \frac{\Sigma a}{n} = M \pm \frac{r}{n}$ in which r is a remainder in division which may be equal to but is generally less than $\frac{n}{2}$ and is rejected.

Now, from the law of the mean the algebraic sum of the departures $=0$,

$$\therefore \Sigma b - tM \mp \frac{r}{n}t + \Sigma c - uM \mp \frac{r}{n}u = 0 \tag{4}$$

$$\text{Let } \Sigma b - tM = B \text{ and } \Sigma c - uM = -C.$$

$$\therefore B - C \mp \frac{r}{n}(t+u) = 0 \tag{5}$$

Since $t+u=n$, $B-C=\pm r$, an equation which checks the accuracy of calculations when B and C are calculated separately.

Changing the signs of the quantities composing the negative departures in equation (4) and substituting B and C, we get for the arithmetical sum of departures,

$$\Sigma V = B + C \mp \frac{r}{n}(u-t)$$

which gives rigorously the sum of departures. Since $\frac{n}{2}$ is a maximum possible value of r and since $(u-t)$ tends to be zero or only small integral numbers, the term $\frac{r}{n}(u-t)$ is a small corrective term which can in general be wholly neglected. $\therefore$

$$\Sigma V = B + C$$

The practical meaning of all this may be stated in a simple rule for computing any values of ΣV, thus:

If not already known, compute the mean M of all the values, noting the amount of any discarded remainder r. Form the sum of all the values of the variant which are *greater* than the mean. Let the number be t. Subtract from the sum, t times the value of the mean. The difference is the sum of the positive departures, B. The sum of the negative departures is $C = B \mp r. \therefore \Sigma V = 2B \mp r$.

If desired, the work can be checked by computing C independently from the sum of the values of *a equal to or less* than the mean. It will be noted for check purposes we also have the relation

$$C = \Sigma a - \Sigma b - uM$$

Formulae similar to the foregoing could be developed for accomplishing the same results by the use of some arbitrary number R instead of the mean M. It is believed the abridgment of the work secured by use of the mean would be lost in the more complex calculations required if an arbitrary number is used.

A single example illustrates the comparative simplicity of this method of computing the mean departure of the results of a *tabulation*. The data are the monthly sums of a 16-year record of rainfall. We get the mean monthly departure without computing either the departures or the monthly means and the result is almost rigorously accurate—that is, fractional excesses are rejected only at the end and are a minimum. The tabulation in figure 3 shows the work as carried out on a listing machine which greatly facilitates the computation.

ABRIDGED CALCULATION OF MEAN DEPARTURE

	Monthly Sums	Sums greater than Mean
$n = 12$	27.32	44.51
	22.57	70.22
	33.56	53.25
	44.51	52.81
	70.22	51.81
	53.25	47.62
	52.81	$320.22 = \Sigma b$
	51.81	$243.24 = tM = uM$
	47.62	$+76.98 = B$
	33.94	$-77.01 = C$ Diff $= -3 = r$
	25.28	$153.99 = \Sigma V$
	23.56	

$6 = t = u$

Sum $= \Sigma a = 486.45$

Mean $= M = 40.54$

$r = -3$

Σa repeated $= 486.45$ $n = 12$

Σb " $= 320.22$ $p = 16$

Diff. $= \Sigma c = 166.23$ $np = 192$

$uM = 243.24$

Diff. $= C = -77.01$

$$V_n = \frac{\Sigma V}{np} = \pm .802$$

FIG. 3.

The mean departure, however, computed, is probably the best and most easily evaluated measure or index of variation we can employ when such a measure is needed. This is especially the case when there are several secondary maxima and minima in a small group of values.

Of course, when periodicities are very elemental and closely harmonic in character, the *amplitude* of the various harmonic elements of a complex cycle is an all-sufficient index. On the other hand, when we are dealing with a complex cycle with many inflexions, little significance attaches to the extreme range between the maximum and the minimum values as an index of variation.

THE INHERENT CHARACTERISTICS OF DATA.

All detailed records of meteorological conditions and many like phenomena of solar and terrestrial activities are of an exceedingly complex character.

Real progress in the study and analysis of such data and interrelations thereof is greatly promoted by a recognition of its inherent characteristics and attention to questions of comparability, homogeneity, uniform weight of values, period of time covered, and other such factors.

It is impossible to discuss these questions exhaustively in this note, and attention will be directed to only the following characteristics which are of special significance in the present connection:

(1) Variation or diversity of similar values.
(2) Order of succession and obvious periodicities.
(3) Frequency distribution:
  (a) Elemental.
  (b) Composite.
  (c) Symmetrical, Gaussian and non-Gaussian.
  (d) Skew.

(1) *The variation or the diversity* among a considerable number of assumed similar values of any meteorological element is of great importance when various combinations of values are made and conclusions deduced from the numerical results secured. Every student of the subject knows that so-called normals from short records are good or poor, depending upon how much variation there is in the individual values. Long records, that is, a large number of individual values are necessary to fix normals of tempertaure and rainfall which in general show great variations, whereas shorter records suffice to fix normals of elements like pressure, which exhibits smaller variations. The same elements show systematically much greater variation in one section of the country than in another. Also, during certain months or seasons the variations are greater than in others. These considerations which guide us in fixing our degree of confidence in the values of important *normals* are just as applicable to the averages of a given number of observations for whatever purpose they may be combined as for normals, and thus it follows that in all careful investigations of data *the diversity of the individual values is an inherent characteristic of much importance which must be properly evaluated and reckoned with.* The methods of doing this by proper allowance for strong and weak observational values, or by weights and otherwise, are so fully covered in the textbooks and generally practiced by students in the more exact sciences, that it is needless to go into such details here. Few students of physical meteorology appear to realize the splendid opportunity the enormous body of meteorological statistics offers for a higher order of statistical analysis and discussion than is frequently practiced. The object of the present effort is to secure attention to these important details of investigation and research.

(2) *Order of succession and obvious periodicities.*—The fundamental feature which identifies a periodicity *is the orderly recurrent succession over and over again of identical phase values.* Elastic vibrations and many like physical phenomena exhibit very perfect periodicity even when the amplitude is very, very small. In meteorology and the inexact sciences, examples of periodicity like the diurnal and annual changes in values of temperature, pressure, rainfall, etc., may also be very definite. In these cases, the *length* of the period is invariable and fixed by astronomical causes and relations. All such periodicities in general are well defined and perfectly obvious by mere inspection and in the majority of cases the amplitude of the periodic features is relatively large.

In cases like the succession of HIGHS and LOWS moving eastward in the extratropical latitudes, the interval between events is, very roughly, three to five days, but is extremely variable and at times there seems to be complete interruption or suspension of the orderly succession of events which, however, are resumed again after a short time. These secondary features of the major phenomena of the general circulation of the atmosphere are attended by a whole train of characteristic changes in the values of pressure, temperature, sunshine, cloudiness, precipitation, winds, etc., all of which recur in sequences of highly irregular length and amplitude.

If we think in units of months, years, etc., obscure and indefinite sequences of the same irregular type and character as just described, all tending to semicyclical recurrence, are marked features of any long record which may be investigated.

Figure 4 is fully representative of practically any body of data which may be presented. A very simple change of scale and interval of time between successive values suffices to adapt the diagrams to become representative of a great many records which may be subjects of investigation.

One of the greatest problems in physical meteorology is to formulate bona fide laws of sequence of the semicyclical succession of values such as have just been discussed. Indeed, the problem is broadly general in many branches of the inexact sciences. Many fragmentary solutions or discoveries of alleged cycles have been offered, as a cycle of seven or eight years in temperature or the Brückner cycle of 35 years. An 11-year cyclical correlation of small percentage between tropical temperatures and sunspots is hesitatingly conceded to be of possible reality. Searching criticisms of such claims weakens rather than strenghtens their foundations of proof. In a last analysis the element of pure chance is such a large factor of domination of the events claimed, or, stating the matter in other words, the margin of reality over purely fortuitous recurrence is so small that such claims have no practical forecasting value. No useful margin of successful verification is possible. In the face of such facts one is puzzled to know how much may be accepted as a matter of physical reality and how much should be rejected as only the operations of chance.

Progress in the study of obscure periodicities requires:

(1) That the reality of the results claimed be established as far as possible by the use of more or less rigorous and analytical methods rather than by resort to arbitrary graphical methods often practiced and which too often tend to nurse into realism the creatures of the imagination.

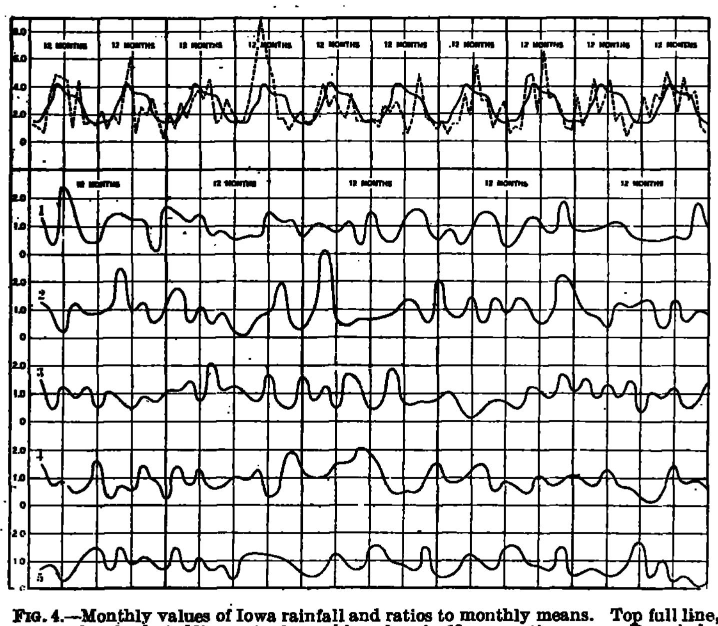

FIG. 4.—Monthly values of Iowa rainfall and ratios to monthly means. Top full line, annual cycle; dotted line, actual monthly values for 10 consecutive years. Remainder of diagram, consecutive values of monthly ratios. Traces 1, 3, and 5 derived from perfectly fortuitous drawings; 2 and 4, actual rainfall data. Real and fortuitous sequences practically indistinguishable.

(2) That the results due exclusively to the operations of chance must be fully evaluated for appropriate comparison with the supposed real results of any investigation in hand. In the attainment of these objects it is of paramount necessity that the original data be reduced to its most elemental form by the complete elimination of all known or obvious periodicities and other characteristic features assignable to some particular cause or associated with a particular time or season. Figure 4 illustrates the results before and after such adjustments have been made on certain rainfall data. As explained in the legend, the results of perfectly fortuitous drawings from the identical body of numbers are also shown in the figure. Without exception it is possible to secure results exactly similar to these, as to actual and fortuitous orders of succession, for any body of data whatever. Such a procedure brings the investigator face to face with the real problem of periodicity. Cyclical sequences derived from the real data that are indistinguishable in their principal features from like sequences deduced from the fortuitous drawings can not be claimed to have physical reality. Demonstrations of reality must be based on results drawn from the real data which can not in any way be duplicated from the fortuitous drawings. It is wonderfully instructive to any investigator to try out and compare for himself the results procurable from drawings, which deals exclusively with the one question of the order of succession of any body of data.

(3) *Frequency distribution—General remarks.*—For purposes of physical investigation of meteorological and like data, frequency distribution must be reduced to the most elemental form possible. The almost universal practice of classifying data on the basis of departure from a mean value is satisfactory only for data like temperature, pressure, etc., where there is no theoretical limit upon the magnitude of either plus or minus departures. Certain classes of data of which rainfall and wind velocity furnish good illustrations have definite zero classes. In such cases values less than zero are hypothetical and impossible. If this kind of data is classified on actual values from zero to the highest, then the resulting distribution for any long record will be highly composite and will most likely tend to be multimodal as shown in figure 5.

Departures from a mean value.—Whether departures from a monthly mean or a mean for the whole group of data are employed a classification by departures, of data having a zero class like rainfall, is unsatisfactory because in either case the zero values of data fall indiscriminately in different places in the frequency distribution.

Ratios to the mean.—A very satisfactory remedy for this difficulty is found by taking the ratio of the individual values of any variant to the mean either of a month, a year, or any other group, or the mean of the whole body of data may be the basis of the ratio. This method brings all zero values into coincidence at zero class and all mean values are coincident at unity or class 1.000. There is no limit to the values which may occur in excess of the mean. The method is essential in classifying data having an absolute zero class, but it is equally advantageous in cases of any other data. The computation of the ratios may seem to require much additional labor, but experience shows not only that subsequent work is simplified by the use of ratios but that such ratios make data, otherwise widely diverse, comparable on a reasonable basis of equality. Regions of light and heavy winds, or of light and heavy precipitation, may be thus readily compared.

Moreover, taking the ratios to the monthly mean effectively eliminates features like the annual cycle. This result is also secured by taking departures from the same means. The latter method, however, introduces a heterogeneous mixture of plus and minus signs which are a fruitful source of errors of computation and are otherwise objectionable.

Figure 6 illustrates a classification of rainfall by ratios, attaining thereby what appears to be a closer approach to the elemental characteristics of the data than is perhaps otherwise possible. The curves and fortuitous drawings in figure 4 are derived from the same data.

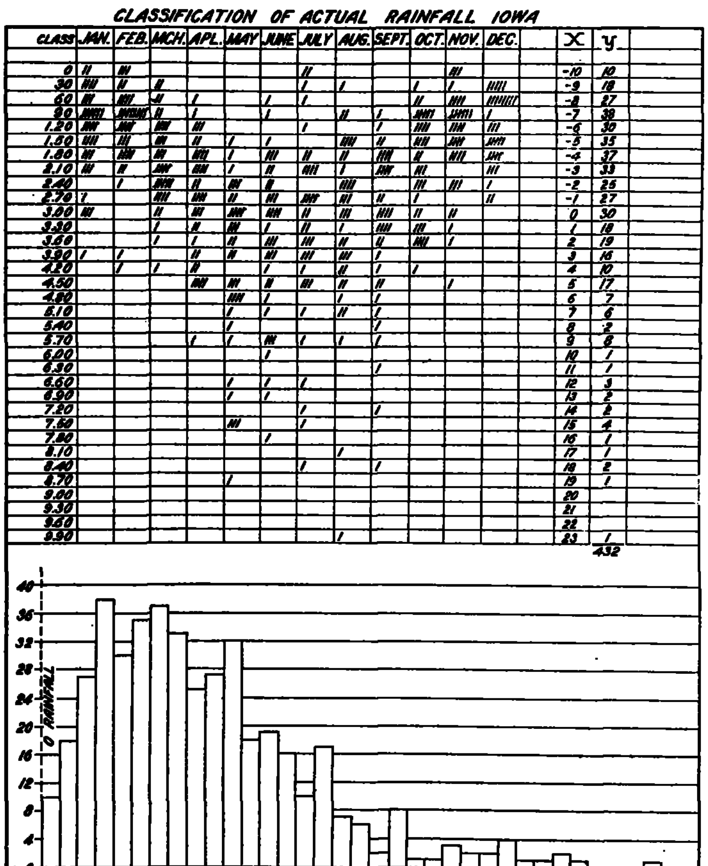

FIG. 5.—Classification and frequency polygon of Iowa rainfall; actual monthly means.

Normal and skew distributions, elemental and composite.—It is a question if any meteorological data exhibit a strictly symmetrical and normal distribution as defined by the Gaussian equations. Variations of temperature when brought into a homogeneous state of values of equal weight and free from diurnal and annual systematic changes appear to be very nearly symmetrical in distribution, but I do not think it has been demonstrated that even these values obey Gauss's law of distribution at all closely.

As a general proposition, all meteorological data form skew distributions well illustrated by diagrams like figures 5 and 6. The Gaussian equations of probabilities apply to these only in the crudest possible way. Even in many cases where the skewness of a distribution is slight, conformity to the Gaussian law is not satisfactory, probably because of some inherent complexity of the data. Take for example a very simple case in which the variations are due to only two separate causes, each of which is Gaussian in its action, we may write from a well-known principle in least squares—

$$\sigma = \pm\sqrt{a^2+b^2}$$

in which a and b are the standard deviations of the separate causes of variation, respectively. Now if the two systems of variations have approximately the same mean value, then the composite distribution will be symmetrical but not elemental, and if a and b differ materially the distribution can not be fitted by a single Gaussian curve with any satisfaction. If, furthermore, the mean values for the two systems differ more or less the resulting distribution will tend to be unsymmetrical and may even have two definite modes.

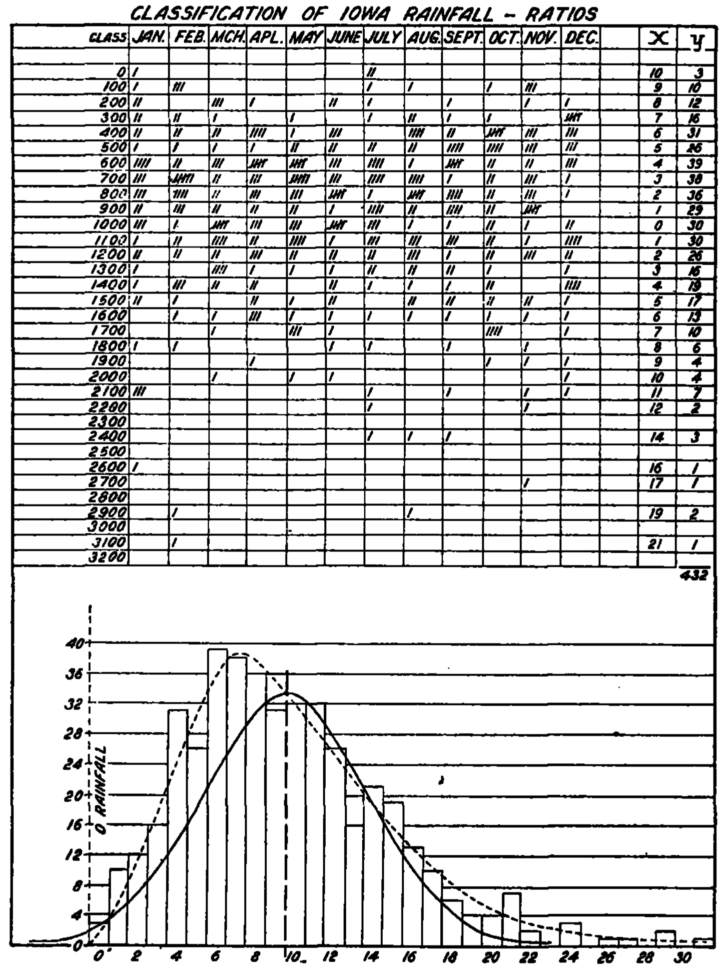

FIG. 6.—Classification and frequency polygon of Iowa rainfall made homogeneous and the annual cycle eliminated by ratios; also Gaussian curve best fit.

Pearson [4] has investigated analytically the case of a composite consisting of two normal distributions and finds its solution both uncertain and very difficult. He finds the result depends upon evaluating the roots of an equation of the ninth degree and selecting those which seem most plausible.

Studies and considerations of this character place the mathematical treatment of composite and unsymmetrical frequency distributions entirely beyond practical possibilities and the grasp of the ordinary investigator. Even the well-known formulæ and methods of Pearson for treatment of certain elemental types of skew distribution are prohibitive because of the great labor of computation entailed, especially when solutions are required for perhaps hundreds of cases. In addition, the result

[4] Pearson, Karl: Contributions to mathematical theory of evolution. *Phil. Trans. Roy. Soc.*, vol. 185, pt. 1, A, 1894, p. 71.

in the end may be disappointing because of the composite character of material handled and the imperfect fit of the theoretical curve to the actual data.

Notwithstanding all such serious obstacles the searching analysis of the frequency distribution of any body of data under discussion is a most essential as well as fruitful source of information, and the case is by no means so hopeless as it seems, because important and very useful phases of the whole problem can be solved most satisfactorily by empirical graphic methods which are ridiculously simple and certain compared with the laborious and possibly disappointing mathematical methods.

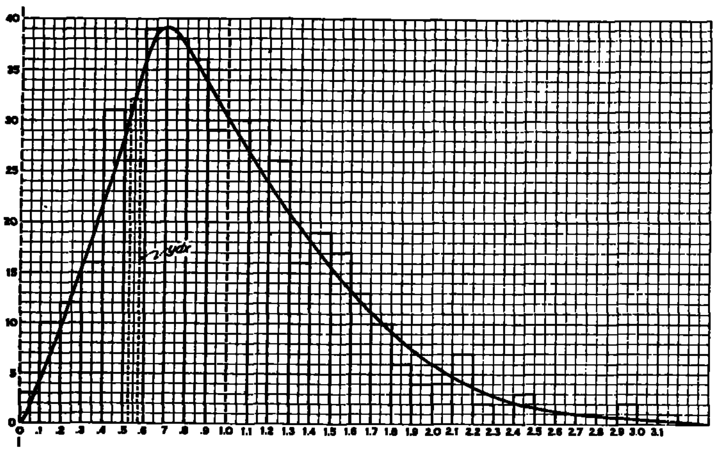

FIG. 7.—Distribution of Iowa rainfall ratios representing graphic integration of skew curve of best fit.

The skew distribution of Iowa rainfall ratios redrawn as figure 7 furnishes a good example, illustrating how the graphic method may be employed.

The more important quantities we wish to determine are:

(1) The modal or most frequent monthly amount of rainfall.

(2) The probability of the modal amount.

(3) The mid or median amount, than which a monthly amount is just as likely to be greater as to be less.

(4) The probability that the monthly rainfall will be the normal for month.

(5) The relations of the mode and median values to the mean.

(6) The probabilities that a departure from the mean will be (*a*) greater than the mean, (*b*) less than the mean.

(7) The most probable positive departure, the one having a 50–50 chance.

(8) The most probable negative departure.

Method.—Draw the frequency distribution carefully to scale on section paper. A size about 10 by 12 inches is ample. Squares about 8 or 10 to the inch are preferable as easy to read. Draw a smooth skew curve avoiding secondary inflections such that the curve incloses an area practically equal to the combined area of the polygon of rectangles.

The answers to all of the above questions could be found (generally, however, with great labor of mathematical computation) *if* the equation of the curve were known and *if* we could integrate the expression ydx. In general, however, no satisfactory equation of the curve can be found, and the integration of ydx is also difficult or impossible. It is perfectly easy, however, to perform a *mechanical integration* of the expression ydx for any free-hand curve, as in figure 7, by simply adding together the coordinates or y values at each intersection of the curve with successive vertical lines of the section ruling. One of the elements ydx is shown in dotted lines in the figure. In this case $dx = 0.05$. This operation of summation is very greatly facilitated by the use of a listing machine which preserves a record of the successive readings and permits introducing subtotals at certain desirable points, such as (*a*) at an ordinate which divides the area at the left of the mean into two equal halves. The place of this ordinate is estimated approximately by eye and the subtotal introduced as the listing proceeds; (*b*) at or near the modal value; (*c*) at or near the mean value of x; finally, (*d*) a subtotal should be taken at or near an ordinate which divides the area to the right of the mean in two equal parts. The grand total of the whole summation gives the total area of the distribution in units of squares of the coordinate rulings. If the readings of the ordinates are estimated to the tenth of a unit, the size and scale of diagram recommended will give a total area of about 1,000 squares, with the nearest tenth added, which is abundantly accurate.

Let A = total area of the distribution. (If this area is not quite equal to the total area (expressed in squares) of the polygons themselves, the curve may need to be adjusted in places to include better the desired area.)

Aided by the subtotals a, b, c, d, it is very easy to interpolate ordinates which accurately fix the following areas:

$$a' = \frac{\text{area from zero to mean}}{2}$$

The abscissa for the ordinate defining this area is the most probable negative departure (a 50–50 chance). Question (8).

The subtotal $b' = \frac{A}{2}$ defines the mid or median ordinate which divides the distribution into two equal parts. Question (3).

The subtotal c' defines the area from 0 to the mean or normal axis, viz, abscissa = 1.00

Subtotal d' locates

$$\frac{\text{area from mean to positive limit of ratios}}{2}$$

This defines the most probable value of rainfall ratio *greater* than the normal, the one having a 50–50 chance of occurring. Question (7).

From the data thus secured numerical answers are easily found for the Iowa rainfall data for 36 years represented by the distribution.

(1) The most frequent (the modal) rainfall we see by inspection is very approximately at $x = .700$; that is, 70 per cent of the monthly normal is the most probable value of monthly rainfall.

(2) The probability that the rainfall in any month will be the modal amount within a narrow limit, say between ratios 0.675 and 0.725 will be $858.6 \div 39$ = about once in 22 times—that is, although the modal rainfall is most frequent of any it will occur as a monthly amount within these limits only about once in 22 months.

(3) The mid or median value of rainfall ratio by interpolation from subtotal c' is 0.917—that is, it is an even chance that any monthly rainfall for this section of Iowa will be greater or less than 92 per cent of the monthly normal.

(4) The frequency for a mean monthly rainfall (ratio 1.00) from the curve is 30.3. The probability that the amount of a monthly rainfall will occur between a ratio say 0.995 and 1.025 will be $858.6 \div 30.3 = 28$—that is, the monthly mean rainfall within ± 0.025 will occur only once in about 28 months.

(5) We have seen *the mode falls at* $x=0.700$, *the* median at 0.917, and the mean or normal at 1.00.

(6) The probability that a monthly value of rainfall will be greater than the normal is measured by the ratio

$$\frac{\text{area greater than mean}}{\text{whole area}(=A)}=\frac{858.6-481.9}{858.6}=0.44$$

Hence the monthly rainfall will equal or be greater than the normal about 44 months in 100 and of course will equal or be less than the normal 56 months.

(7) For monthly amounts greater than the normal the percentage 1.36 is the probable amount.

(8) *If a monthly rainfall is less than the normal it* will be an even chance that the amount will be greater or less than 66 per cent of the normal, and it was shown under (1) that the most frequent of all monthly amounts was 70 per cent of the normal. Thus it appears that the most frequent monthly amounts and the probable amounts below the average are both about two-thirds the monthly normal.

Such are answers that are easily deduced by interpolations from the mechanical or approximate integration of such scale drawings of frequency distributions as shown in figure 7.

A STATISTICAL COMPARISON OF METEOROLOGICAL DATA WITH DATA OF RANDOM OCCURRENCE.

551.506 : 551.501

By H. W. CLOUGH.

[Weather Bureau, Washington, D. C., Apr. 18, 1921.]

SYNOPSIS.

Daily, monthly and annual means of meteorological data show fluctuations of varying orders of magnitude, which may be regarded as either of a fortuitous character or as presenting more or less systematic characteristics. Certain precise relations which are distinctive of purely fortuitous data are derived by both theoretical and empirical methods. These relations constitute criteria for determining the extent to which meteorological data differ from such fortuitous data.

Monthly and annual means of temperature are nearly Gaussian in their distribution, their deviations being of the nature of accidental errors, but the order of succession of their occurrence is not fortuitous. Rainfall data are more fortuitous in their characteristics than temperature. In a plot of unrelated numbers the two-interval is predominant, while in the case of most meteorological annual means the three-year interval is the most frequent. The variations of mean annual temperatures show systematic characteristics to a greater extent in the Southern Hemisphere and the low latitudes of the Northern Hemisphere than in the higher latitudes of the Northern Hemisphere.

Statistical criteria applied to the variations of the period of the solar spots disclose markedly systematic characteristics.

A period of recurrence of extremes of pressure at Toronto, averaging 32 to 34 days seems to be disclosed by a purely statistical method of treatment of the dates of highest and lowest pressure in each month for a long series of years.

Variability is a dominant characteristic of weather, particularly in temperate latitudes. In the Tropics the day-to-day fluctuations are negligible and the seasonal changes occur with clock-like regularity. The interdiurnal variability of temperature increases with latitude to about the Arctic Circle, then decreases somewhat. A plot of daily mean temperatures exhibits characteristic fluctuations with crests separated by intervals varying irregularly from 3 to 7 days or more. If these daily values be combined into weekly means and plotted there are again shown similar fluctuations but with longer intervals varying from 2 to 5 or 6 weeks. The same data combined into monthly means show, when the residuals are plotted, fluctuations apparently analogous to those of the daily data but with intervals between the successive crests varying from 2 to 6 months or more. Yearly mean temperatures at any locality when plotted show fluctuations which are indistinguishable from a plot of monthly residuals, the intervals being measured in years instead of months.

Thus daily, weekly, monthly, and annual means of meteorological data present fluctuations of varying orders of magnitude. The smaller day-to-day fluctuations are superposed upon the larger weekly fluctuations, the weekly upon the larger monthly, and so on until we arrive at the long secular variations measured by decades or even centuries.

The question arises as to the character of these apparently irregular fluctuations. Are they to be regarded as purely accidental and fortuitous or do they present characteristics which show them to be deviations partaking of a systematic nature and if so, susceptible of prediction? Obviously, if they are of a purely fortuitous nature long-range forecasting is out of the question.

Considerable diversity of opinion regarding this particular aspect of weather changes is gleaned from the literature of the subject. A conservative element regards the monthly, seasonal, and annual variations as due to a complex set of many varying influences whose resultant effect is a series of nearly fortuitous deviations about the normal which can be represented by the well-known Gaussian law of errors. Another element regards the variations as controlled by more or less systematic laws and as being essentially sequences of a quasi-periodic nature. Popular weather lore has for its basis an almost universal belief in the tendency of weather changes to be complementary, in other words for one extreme to be followed by the opposite within a short period.

Obviously it is possible by the employment of statistical criteria to determine the extent to which a given succession of meteorological data conforms to a purely fortuitous selection of similar data, and it will be the purpose of this paper to set forth the characteristics of data which represent purely accidental deviations about a mean and to illustrate by examples of meteorological data how and to what extent the latter differ from data of random occurrence.

CHARACTERISTIC FEATURES OF FORTUITOUS DATA.

There are two classes of data whose deviations present purely fortuitous characteristics: (1) A series of unrelated numbers, illustrated by a random selection of numbers between 0 and 100. In this class of data all values are equally probable. (2) A sample of the component data of a normal frequency distribution, illustrated by the sums of ten digits of random selection. The data of this class are also unrelated, but the various possible values of the variant are of unequal probability.

There are certain precise relations to which these two classes of data rigorously conform and which constitute criteria for testing the conformity of any series of observational data to these requirements. Any deviation from such a conformity indicates some systematic influence operating which results in a frequency curve of either a skew type or a symmetrical but composite type. In either case the curve of best fit by least square methods exhibits excesses in one part of the curve and deficiencies in another part.

Relations between indices of dispersion.—There are various measures of the dispersion or scatter, which, in the case of data showing a distribution of a